ESTRATEGIAS Y ANÁLISIS ORIENTADOS AL MANEJO DE DATOS MASIVOS USANDO COMPUTACIÓN DE ALTO DESEMPEÑO

Mercedes Barrionuevo, Mariela Lopresti, Maximiliano Lucero, Natalia Miranda, Cristian Pérez Monte, M. Antonia Murazzo(*), Fabiana Píccoli y Marcela Printista.

LIDIC- Univ. Nacional de San Luís

(*) Univ. Nacional de San Juan

San Luís, Argentina

{mdbarrio, omlopres, mlucero, ncmiran, mpiccoli, mprinti}@unsl.edu.ar

RESUMEN

Los datos en nuestro universo digital han crecido de tera bytes a zetta bytes. Pero no todos los datos existentes son significativos. Un gran desafío para muchos investigadores es el descubrimiento de conocimiento a partir de un conjunto de datos muy grande en un tiempo razonable. Para lograrlo hoy se piensa en arquitecturas de naturaleza heterogéneas formadas por procesadores many y multicores.

En este trabajo se expone distintas líneas de trabajo a seguir teniendo como objetivo desarrollar técnicas de Computación de Alto Desempeño para resolver este tipo de problemas.

Palabras clave: Big Data. Computación de Alto Desempeño. GPUs. Tráfico en Redes de Computadoras**.**

CONTEXTO

Esta propuesta de trabajo se lleva a cabo dentro del proyecto de investigación "*Tecnologías Avanzadas aplicadas al Procesamiento de Datos Masivos*" y del proyecto binacional CAPG-BA 66/13 entre la Universidad Nacional de San Luis y la Universidad de Pernambuco, Recife, Brasil.

El proyecto de investigación se desarrolla en el marco del Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC), de la Facultad de Ciencias Físico, Matemáticas y Naturales de la Universidad Nacional de San Luis y el Centro de Informática de la UFPE.

1. INTRODUCCIÓN

El crecimiento de datos circulantes en Internet, las nuevas tecnologías y el aumento en la velocidad de transmisión de datos han dado origen al concepto de Big Data o Datos Masivos [MCJ13]. Este concepto, actualmente se define haciendo referencia a siete características: Variedad, Volumen, Velocidad, Veracidad, Viabilidad, Visualización y Valor [MCFEE12].

El conjunto de datos a manipular es tan grande y complejo que los medios tradicionales de procesamiento son ineficaces. Por lo cual es un desafío analizar, capturar, recolectar, buscar, compartir, almacenar, transferir, visualizar, etc., cantidades masivas de información, obtener conocimiento y

realizar toda su gestión en un tiempo razonable [N13].

Existen diferentes áreas o aplicaciones donde se trabaja con gran cantidad de datos, entre ellas podemos destacar la seguridad en redes de datos, recuperación de la información, evolución y parentesco de las especies, entre otras.

De acuerdo a todo lo expuesto, el procesamiento de grandes volúmenes de datos hacen que los sistemas de cómputo convencionales sean muchas veces inapropiados para lograr un procesamiento adecuado, por lo tanto una alternativa es considerar técnicas de Computación de Alto Desempeño (HPC) [YOU14], las cuales permiten realizar operaciones de cómputo intensivo y mejorar la velocidad de procesamiento; involucrando diferentes tecnologías tal como los sistemas distribuidos y los sistemas paralelos (cluster de computadoras, cloud computing, tarjetas gráficas y computadoras masivamente paralelas) [HAGER10].

Además de las técnicas de HPC, es importante la arquitectura subyacente. Hoy en día, la evolución de los sistemas de computación con multiprocesadores ha seguido dos líneas de desarrollo: las arquitecturas multicore (multi-núcleos) y las arquitecturas manycores (muchos-núcleos) [HWU08], ambos ofrecen un acceso rápido a una única memoria compartida, evitando la transferencia de datos entre distintas máquinas a través de una red. Sin embargo la cantidad de memoria disponible es limitada, una alternativa son las arquitecturas con memoria distribuida, las cuales permiten incrementar el espacio de almacenamiento (principal y secundario) aunque deben pagar el precio de la latencia de la red para llevar a cabo las comunicaciones [KAUR14]. Otra alternativa son los sistemas híbridos, los cuales permiten combinar las características de sistemas con memoria compartida y memoria distribuida, multicores con GPU, varios sistemas en la red, entre otros. De esta manera es posible incrementar la capacidad y poder de cómputo de los sistemas computacionales permitiendo la ejecución en paralelo de múltiples procesos y threads con distintas administraciones de memoria.

Todo lo expuesto anteriormente constituye la base de nuestra motivación para investigar, verificar y poner en marcha nuevas técnicas y arquitecturas que ayuden a mejorar el procesamiento y sus tiempos de respuesta. Las técnicas de HPC serán nuestras herramientas para resolver con eficiencia cada uno de los objetivos: aplicar técnicas de HPC adecuadas para resolver problemas de datos masivos en ambientes paralelos híbridos en arquitecturas many y multicore. Como caso de estudio se presenta la búsqueda de soluciones a los diferentes los problemas planteados en la siguiente sección.

2. LÍNEAS DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN

Diferentes áreas de la ciencia y la sociedad deben considerarse como casos de problemas Big Data. Estas áreas constituyen sendas líneas de investigación descriptas a continuación:

- *Análisis de tráfico en términos de la seguridad de la información:*

  En el campo de la detección de anomalías en redes de datos, el problema consiste en la identificación de patrones no acordes al comportamiento normal del tráfico en la red [BLMP16]. Detectar un posible ataque requiere contar con tecnologías para la clasificación del tráfico, asociando flujos de datos con las

aplicaciones que los generan. El conjunto de datos con los cuales se trabaja crece a gran velocidad, mucho mayor que su capacidad de procesamiento.

Esta línea tiene como objetivo la creación de modelos para la búsqueda de anomalías, y más precisamente, de ataques a redes de datos. Esta búsqueda consiste en la identificación de patrones que se desvían del comportamiento normal de tráfico. Tener la capacidad de tratar grandes volúmenes de información no solamente nos permite saber qué es lo que está pasando en este instante, sino también trazar patrones a lo largo del tiempo. Muchas veces es fácil pasar por alto algunos indicadores cuando analizamos información en tiempo real, sin embargo, si analizamos esa información en otros contextos y a lo largo del tiempo, quizá podamos encontrar otros significados. Lo que se pretende realizar es el procesamiento del tráfico de red, aplicarle inteligencia para obtener resultados concretos y en un tiempo cercano, de forma que la conclusión arribada no sea irrelevante, constituya la detección de un ataque y se puedan tomar decisiones rápidas.

- *Recuperación de la Información usando índices de búsquedas:*

  En el caso de la recuperación de la información, el objetivo principal es satisfacer la necesidad de respuestas planteadas por un usuario en lenguaje natural, especificada a través de un conjunto de palabras claves. La información multimedia debe ser recuperada aplicando búsquedas por similitud, por lo cual se necesitan métodos de acceso eficientes que permitan recuperar rápidamente los elementos que satisfacen los criterios de consulta. Un sistema de recuperación de información encuentra datos importantes con coincidencia parcial al patrón dado. Esta similitud, se modela utilizando una función de distancia, la cual satisface entre otras propiedades la desigualdad triangular.

  Para realizar consultas en bases de datos no estructuradas o métricas se han creado modelos y algoritmos de búsqueda más generales que los correspondientes a bases de datos tradicionales [CNBY01]. El tipo de consulta en estas bases de datos se denominan consultas por similitud o proximidad, aquí los elementos son buscados considerando la cercanía de los mismos al elemento consultado. El objetivo de esta línea es utilizar índices de búsquedas aproximadas como los *permutantes* [LMPR13] para resolver consultas.

- *Construcción de Árboles Filogenéticos*:

  En el área de biología evolutiva [B09, G14, Z14], el estudio de las relaciones evolutivas entre especies a partir de la distribución de los caracteres primitivos, utilizando información ADN y de morfología, es una tarea que se encuadra dentro del área de datos masivos. La idea es elaborar un árbol filogenético en el cual se muestran las relaciones evolutivas de los seres vivos entre sí, es decir, reconocer los grados de cercanía de

ancestros comunes. Este trabajo requiere contar con datos moleculares (fundamentalmente ADN) y/o morfológicos, los cuales identificarán la/s especie/s a incorporar en el árbol. Este proceso demanda gran cantidad de recursos computacionales.

El análisis filogenético requiere cálculos más complejos a medida que la cantidad de especies a catalogar crece, la cantidad de árboles a examinar en una búsqueda aumenta exponencialmente. Por esta razón es necesario contar con métodos, herramientas y técnicas de computación de altas prestaciones para datos masivos con el objeto de obtener información dentro de tiempos razonables.

Todas estas líneas de investigación tienen en cuenta la portabilidad de los desarrollos, esto se debe a que las soluciones a plantear tienen como objetivo trabajar en arquitecturas heterogéneas, aplicando técnicas de paralelismo híbrido y logrando tiempo de respuestas menores a cada uno de los problemas planteados.

3. RESULTADOS OBTENIDOS/ESPERADOS

Como objetivos de las líneas de investigación nos planteamos facilitar el desarrollo de soluciones paralelas portables, de costo predecible, capaces de explotar las ventajas de modernos ambientes de HPC a través de herramientas y "frameworks de computación" de alto nivel. Para ello será necesario proponer nuevas metodologías a ser aplicadas en cada una de las fases del tratamiento de datos masivos.

4. FORMACIÓN DE RECURSOS HUMANOS

Los resultados esperados respecto a la formación de recursos humanos son hasta el momento el desarrollo de 4 tesis doctorales (2 concluidas), 2 tesis de maestría y varias tesinas de grado.

5. BIBLIOGRAFIA

[B09] A. Benítez Burraco, *"Genes y lenguaje: aspectos ontogenéticos, filogenéticos y cognitivos"*. ISBN 8429110046, 9788429110043. Reverte, 2009.

[BLMP16] Mercedes Barrionuevo, Mariela Lopresti, Natalia Miranda, Fabiana Píccoli. "*Un enfoque para la detección de anomalías en el tráfico de red usando imágenes y técnicas de Computación de Alto Desempeño*". XXII Congreso Argentino De Ciencias de la Computación. CACIC 2016. Pp. 1166-1175. Octubre 2016, San Luis, Argentina.

[CNBY01] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. 2001. Searching in metric spaces. *ACM Comput. Surv.* 33, 3 (September 2001), 273-321.

[G14] L. Garamszegi. "*Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice"*. ISBN 3662435500, 9783662435502. Springer, 2014.

[HAGER10] Hager, Georg and Wellein, Gerhard. "*Introduction to high performance computing for scientists and engineers*". Chapman & Hall/CRC computational science

series. 1st edition, 2010. ISBN 978-1-4398-1192-4.

[HWU08] W. Hwu, K. Keutzer, and T. G. Mattson, "*The Concurrency Challenge*" IEEE Des. Test Comput., vol. 25, no. 4, pp. 312–320, Jul. 2008.

[KAUR14] K. Kaur and A. K. Rai, "A Comparative Analysis: Grid, Cluster and Cloud Computing," Int. J. Adv. Res. Comput. Commun. Eng., vol. 3, no. 3, pp. 2278–1021, 2014

[LMPR13] Lopresti, M., Miranda, N., Piccoli, F., Reyes, N. - Solving Multiple Queries through the Permutation Index in GPU. 4th International supercomputing Conference in Mexico. Colima-México. 5-8 Marzo 2013.

[MCFEE12] A. Mcafee, E. Brynjolfsson, and H. Org, "*Big Data: The Management Revolution SPOTLIGHT ON BIG DATA*" 2012.

[MCJ13] V. Mayer-Schönberger, K. Cukier. A.I. Jurado. "*Big data: La revolución de los datos masivos*". Turner. 2013.

[N13] J. Needham. "*Disruptive Possibilities: How Big Data Changes Everything*". Kindle Edition. O'Reilly Media Inc. 2013.

[YOU14] Y. You, S. L. Song, H. Fu, A. Marquez, M. M. Dehnavi, K. Barker, K. W. Cameron, A. P. Randles, and G. Yang, "*MIC-SVM: Designing a Highly Efficient Support Vector Machine for Advanced Modern Multi-core and Many-Core Architectures,*" in 2014 IEEE 28th International Parallel and Distributed Processing Symposium, 2014, pp. 809–818.

[Z14] Z. Yang, *"Molecular Evolution: A Statistical Approach"*. OUP Oxford, 2014.